Load and Clean datasets

Aleksei Sorokin, asorokin@hawk.iit.edu, A20394300

3/3/2020

Notes

Data sourced from https://www.sports-reference.com/

Packages

```
library(stringr)
library(rvest)
library(tidyr)
```

Scrape & clean football coaches

```
url_fb <- 'https://www.pro-football-reference.com/coaches/'
t_fb <- html_nodes(read_html(url_fb), css = 'table')
df_fb_og <- html_table(t_fb[[1]])
# rename original columns
colnames(df_fb_og) = c(
  'RK', # rank
  'N', # coach name
  'Y', # total years coaching
  'YR', # range of years coached in
  'G', # total games coached
  'W', # total wins
  'L', # total losses
  'T', # ties
  'WL%', # win-loss %
  'GO500', # number of games over .500 (wins-losses)
  'PY', # playoff years
  'PG', # years coach made playoffs
  'PW', # playoff wins
  'PL', # playoff losses
  'PWL%', # playoff win-loss %
  'MCR', # mean conference rank (16 teams per conference)
  'BCR', # best conference rank
  'C', # championships (includes super bowls or championships)
  'SBW', # super bowl champions
  'CC') # conference championships
# drop useless and dependent columns
drop_cols <- c(
  'RK', # don't need rank
  'W','L','T','GO500', # total games and win-loss % are sufficient
  'PW','PL', # playoff games and playoff win-loss % are sufficient
  'PY','MCR','BCR', # not consistently provided across all datasets
  'SBW') # championships includes super-bowl wins and championships before super-bowl
df_fb <- df_fb_og[,!(names(df_fb_og)%in%drop_cols)]
# set na values to 0
```

```
df_fb[is.na(df_fb)] <- 0
# extract hall of fame indicator (1=HOF, 0=!HOF)
df_fb$HOF <- grepl('\\+',df_fb[['N']])
df_fb$HOF <- as.numeric(df_fb$HOF)
# clean up names
df_fb$N <- gsub('\\+','',df_fb[['N']])
df_fb$N <- str_squish(df_fb$N)
# split year range and only keep final year coaching
df_fb[,c('YS','YE')] <- do.call(rbind,strsplit(df_fb$YR,'-'))
df_fb$YE <- as.numeric(df_fb$YE)
df_fb <- df_fb[,!(names(df_fb)%in%c('YS','YR'))] # drop year range for year end
# ensure numeric datatypes
for (col in (names(df_fb))){if (col !='N'){df_fb[,col] <- as.numeric(df_fb[,col])}}
# reindex
rownames(df_fb) = 1:nrow(df_fb)
# output
head(df_fb)
```

```
##                N  Y   G   WL% PG  PWL% C CC HOF   YE
## 1      Don Shula 33 490 0.677 36 0.528 2  6   1 1995
## 2   George Halas 40 497 0.682  9 0.667 6  0   1 1967
## 3 Bill Belichick 25 400 0.683 43 0.721 6  9   0 2019
## 4     Tom Landry 29 418 0.607 36 0.556 2  5   1 1988
## 5  Curly Lambeau 33 380 0.631  5 0.600 6  0   1 1953
## 6     Paul Brown 25 326 0.672 17 0.529 7  0   1 1975
```

```
summary(df_fb)
```

```
##       N                   Y                G               WL%
##  Length:500         Min.   : 1.000   Min.   :  1.00   Min.   :0.0000
##  Class :character   1st Qu.: 1.000   1st Qu.: 12.00   1st Qu.:0.2650
##  Mode  :character   Median : 3.000   Median : 32.00   Median :0.4150
##                     Mean   : 4.768   Mean   : 64.13   Mean   :0.3961
##                     3rd Qu.: 6.000   3rd Qu.: 83.25   3rd Qu.:0.5320
##                     Max.   :40.000   Max.   :497.00   Max.   :1.0000
##        PG              PWL%              C                CC
##  Min.   : 0.000   Min.   :0.0000   Min.   :0.000   Min.   :0.000
##  1st Qu.: 0.000   1st Qu.:0.0000   1st Qu.:0.000   1st Qu.:0.000
##  Median : 0.000   Median :0.0000   Median :0.000   Median :0.000
##  Mean   : 2.312   Mean   :0.1342   Mean   :0.224   Mean   :0.216
##  3rd Qu.: 2.000   3rd Qu.:0.2000   3rd Qu.:0.000   3rd Qu.:0.000
##  Max.   :43.000   Max.   :1.0000   Max.   :7.000   Max.   :9.000
##       HOF              YE
##  Min.   :0.000   Min.   :1920
##  1st Qu.:0.000   1st Qu.:1942
##  Median :0.000   Median :1974
##  Mean   :0.126   Mean   :1971
##  3rd Qu.:0.000   3rd Qu.:2003
##  Max.   :1.000   Max.   :2019
```

Scrape & clean baksetball coaches

```
url_bkb <- 'https://www.basketball-reference.com/coaches/NBA_stats.html'
t_bkb <- html_nodes(read_html(url_bkb), css = 'table')
df_bkb_og <- html_table(t_bkb[[1]])[-1,]
# rename original columns
colnames(df_bkb_og) = c(
  'RK', # rank
  'N', # coach name
  'YS', # first year coaching
```

```
  'YE', # last year coaching
  'Y', # total years coaching
  'G', # total games coached
  'W', # total wins
  'L', # total losses
  'WL%', # win-loss %
  'GO500over2', # number of games over .500 (wins-losses)/2
  'PG', # years coach made playoffs
  'PW', # playoff wins
  'PL', # playoff losses
  'PWL%', # playoff win-loss %
  'CC', # conference championships
  'C') # championships
# remove header rows
df_bkb <- df_bkb_og[!(df_bkb_og$G=='Regular Season' | df_bkb_og$G=='G'),]
# drop useless and dependent columns
drop_cols <- c(
  'RK', # don't need rank
  'YS', # captured by total years (Y) and last year coaching (YE)
  'W','L','GO500over2', # total games and win-loss % are sufficient
  'PW','PL') # playoff games and playoff win-loss % are sufficient
df_bkb <- df_bkb[,!(names(df_bkb)%in%drop_cols)]
# set columns to be numeric
for (col in (names(df_bkb))){if (col !='N'){df_bkb[,col] <- as.numeric(df_bkb[,col])}}
# set na values to 0
df_bkb[is.na(df_bkb)] <- 0
# extract hall of fame indicator (1=HOF, 0=!HOF)
df_bkb$HOF <- grepl('\\*',df_bkb[['N']])
df_bkb$HOF <- as.numeric(df_bkb$HOF)
# clean up names
df_bkb$N <- gsub('\\*','',df_bkb[['N']])
df_bkb$N <- str_squish(df_bkb$N)
# reindex
rownames(df_bkb) = 1:nrow(df_bkb)
# output
head(df_bkb)
```

```
##                 N   YE  Y    G   WL%  PG  PWL% CC C HOF
## 1    Rick Adelman 2014 23 1791 0.582 157 0.503  2 0   0
## 2  Richie Adubato 1997  6  367 0.346   8 0.250  0 0   0
## 3     Danny Ainge 2000  4  226 0.602  12 0.250  0 0   0
## 4     Stan Albeck 1986  7  574 0.535  44 0.409  0 0   0
## 5 Curly Armstrong 1949  1   54 0.407   0 0.000  0 0   0
## 6  Kenny Atkinson 2020  4  308 0.383   5 0.200  0 0   0
```

```
summary(df_bkb)
```

```
##       N                   YE             Y                G
##  Length:332         Min.   :1947   Min.   : 1.000   Min.   :   1.0
##  Class :character   1st Qu.:1972   1st Qu.: 1.000   1st Qu.:  60.0
##  Mode  :character   Median :1996   Median : 3.000   Median : 169.5
##                     Mean   :1990   Mean   : 5.361   Mean   : 368.3
##                     3rd Qu.:2011   3rd Qu.: 7.000   3rd Qu.: 534.2
##                     Max.   :2020   Max.   :32.000   Max.   :2487.0
##       WL%               PG              PWL%              CC
##  Min.   :0.0000   Min.   :  0.00   Min.   :0.0000   Min.   : 0.0000
##  1st Qu.:0.3237   1st Qu.:  0.00   1st Qu.:0.0000   1st Qu.: 0.0000
##  Median :0.4260   Median :  3.00   Median :0.0000   Median : 0.0000
##  Mean   :0.4146   Mean   : 24.17   Mean   :0.2154   Mean   : 0.2952
##  3rd Qu.:0.5250   3rd Qu.: 27.00   3rd Qu.:0.4313   3rd Qu.: 0.0000
##  Max.   :0.7120   Max.   :333.00   Max.   :0.7330   Max.   :13.0000
##        C                HOF
```

```
##  Min.   : 0.0000   Min.   :0.00000
##  1st Qu.: 0.0000   1st Qu.:0.00000
##  Median : 0.0000   Median :0.00000
##  Mean   : 0.2199   Mean   :0.06627
##  3rd Qu.: 0.0000   3rd Qu.:0.00000
##  Max.   :11.0000   Max.   :1.00000
```

Scrape and clean baseball coaches

```
url_bb <- 'https://www.baseball-reference.com/managers/'
t_bb <- html_nodes(read_html(url_bb), css = 'table')
df_bb_og <- html_table(t_bb[[1]])
# rename original columns
colnames(df_bb_og) = c(
  'RK', # rank
  'N', # coach name
  'Y', # total years coaching
  'YS', # first year coaching
  'YE', # last year coaching
  'W', # total wins
  'L', # total losses
  'WL%', # win-loss %
  'T', # ties
  'GO500', # games over 500 (W-L)
  'G', # total games coached
  'PW', # playoff wins
  'PL', # playoff losses
  'PWL%', # playoff win-loss %
  'BF', # best finish
  'WF', # worst finish
  'MRK', # mean rank
  'E', # ejections
  'PY', # years in the playoffs
  'C',# championships (World Series wins)
  'CC', # conference championships (pennant wins)
  'AGM', # all star games managed
  'PS', # player stats
  'PMY') # years as player or manager
# remove header rows
df_bb <- df_bb_og[!(df_bb_og$N=='Mgr'),]
# set columns to be numeric
for (col in (names(df_bb))){if (col !='N'){df_bb[,col] <- as.numeric(df_bb[,col])}}
```

```
## Warning: NAs introduced by coercion
```

```
## Warning: NAs introduced by coercion
```

```
# calculate PG (postseason games)
df_bb$PG <- df_bb$PW+df_bb$PL
# drop useless and dependent columns
drop_cols <- c(
  'RK', # don't need rank
  'YS', # captured by total years (Y) and last year coaching (YE)
  'W','L','T','GO500', # total games and win-loss % are sufficient
  'OL','PTS','PTS%',  # not consistently provided across all datasets
  'PW','PL', # playoff games and playoff win-loss % are sufficient
  'BF','WF','MRK','E','PY','AGM','PS','PMY') # not consistently provided across all datasets
df_bb <- df_bb[,!(names(df_bb)%in%drop_cols)]
# set na values to 0
df_bb[is.na(df_bb)] <- 0
# extract hall of fame indicator (1=HOF, 0=!HOF)
```

```
df_bb$HOF <- grepl('HOF',df_bb[['N']])
df_bb$HOF <- as.numeric(df_bb$HOF)
# clean up names
df_bb$N <- gsub('HOF','',df_bb[['N']])
df_bb$N <- str_squish(df_bb$N)
# reindex
rownames(df_bb) = 1:nrow(df_bb)
# output
head(df_bb)
```

```
##              N Y   YE   WL%   G PWL% C CC PG HOF
## 1   Manny Acta 6 2012 0.418 890    0 0  0  0   0
## 2   Bill Adair 1 1970 0.400  10    0 0  0  0   0
## 3   Joe Adcock 1 1967 0.463 162    0 0  0  0   0
## 4     Bob Addy 2 1877 0.258  31    0 0  0  0   0
## 5    Bob Allen 2 1900 0.500 179    0 0  0  0   0
## 6 Doug Allison 1 1873 0.087  23    0 0  0  0   0
```

```
summary(df_bb)
```

```
##       N                   Y                YE            WL%
##  Length:711         Min.   : 1.000   Min.   :1871   Min.   :0.0000
##  Class :character   1st Qu.: 1.000   1st Qu.:1900   1st Qu.:0.3995
##  Mode  :character   Median : 3.000   Median :1951   Median :0.4720
##                     Mean   : 4.895   Mean   :1948   Mean   :0.4481
##                     3rd Qu.: 6.000   3rd Qu.:1990   3rd Qu.:0.5165
##                     Max.   :53.000   Max.   :2019   Max.   :1.0000
##        G               PWL%               C                CC
##  Min.   :   1.0   Min.   :0.00000   Min.   :0.0000   Min.   : 0.0000
##  1st Qu.:  79.0   1st Qu.:0.00000   1st Qu.:0.0000   1st Qu.: 0.0000
##  Median : 269.0   Median :0.00000   Median :0.0000   Median : 0.0000
##  Mean   : 620.5   Mean   :0.09388   Mean   :0.1688   Mean   : 0.3938
##  3rd Qu.: 780.0   3rd Qu.:0.00000   3rd Qu.:0.0000   3rd Qu.: 0.0000
##  Max.   :7755.0   Max.   :1.00000   Max.   :7.0000   Max.   :10.0000
##        PG               HOF
##  Min.   :  0.00   Min.   :0.0000
##  1st Qu.:  0.00   1st Qu.:0.0000
##  Median :  0.00   Median :0.0000
##  Mean   :  4.54   Mean   :0.1322
##  3rd Qu.:  0.00   3rd Qu.:0.0000
##  Max.   :142.00   Max.   :1.0000
```

Scrape and clean hockey coaches

```
url_h <- 'https://www.hockey-reference.com/coaches/NHL_stats.html'
t_h <- html_nodes(read_html(url_h), css = 'table')
df_h_og <- html_table(t_h[[1]])[-1,]
# rename original columns
colnames(df_h_og) = c(
  'RK', # rank
  'N', # coach name
  'YS', # first year coaching
  'YE', # last year coaching
  'Y', # total years coaching
  'G', # total games coached
  'W', # total wins
  'L', # total losses
  'T', # ties
  'OL', # overtime losses
  'PTS', # points
```

```
  'PTS%', # points / total possible points
  'PG', # years coach made playoffs
  'PW', # playoff wins
  'PL', # playoff losses
  'PT', # playoff ties
  'PWL%', # playoff win-loss %
  'CC', # conference championships
  'C') # championships (Stanley Cup wins)
# remove header rows
df_h <- df_h_og[!(df_h_og$RK=='RK' | df_h_og$N=='Coach'),]
# set columns to be numeric
for (col in (names(df_h))){if (col !='N'){df_h[,col] <- as.numeric(df_h[,col])}}
# calculate wl% (win loss %)
df_h$`WL%` <- df_h$W/(df_h$W+df_h$L)
# drop useless and dependent columns
drop_cols <- c(
  'RK', # don't need rank
  'YS', # captured by total years (Y) and last year coaching (YE)
  'W','L','T', # total games and win-loss % are sufficient
  'OL','PTS','PTS%',  # not consistently provided across all datasets
  'PW','PL','PT') # playoff games and playoff win-loss % are sufficient
df_h <- df_h[,!(names(df_h)%in%drop_cols)]
# set na values to 0
df_h[is.na(df_h)] <- 0
# extract hall of fame indicator (1=HOF, 0=!HOF)
df_h$HOF <- grepl('\\*',df_h[['N']])
df_h$HOF <- as.numeric(df_h$HOF)
# clean up names
df_h$N <- gsub('\\*','',df_h[['N']])
df_h$N <- str_squish(df_h$N)
# reindex
rownames(df_h) = NULL
# output
head(df_h)
```

```
##              N   YE  Y   G  PG  PWL% CC C        WL% HOF
## 1     Sid Abel 1976 16 964  76 0.421  0 0 0.47218789   1
## 2   Jack Adams 1947 20 964 105 0.500  3 3 0.51432130   1
## 3   Gary Agnew 2007  1   5   0 0.000  0 0 0.00000000   0
## 4  Keith Allen 1969  2 150  11 0.273  0 0 0.43220339   0
## 5 Dave Allison 1996  1  25   0 0.000  0 0 0.08333333   0
## 6 Jim Anderson 1975  1  54   0 0.000  0 0 0.08163265   0
```

```
summary(df_h)
```

```
##       N                   YE            Y                G
##  Length:377         Min.   :1919   Min.   : 1.000   Min.   :   1.0
##  Class :character   1st Qu.:1975   1st Qu.: 1.000   1st Qu.:  61.0
##  Mode  :character   Median :1991   Median : 3.000   Median : 163.0
##                     Mean   :1986   Mean   : 4.915   Mean   : 306.7
##                     3rd Qu.:2009   3rd Qu.: 6.000   3rd Qu.: 390.0
##                     Max.   :2020   Max.   :30.000   Max.   :2141.0
##        PG              PWL%             CC               C
##  Min.   :  0.00   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:  0.00   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000
##  Median :  6.00   Median :0.2500   Median :0.0000   Median :0.0000
##  Mean   : 23.44   Mean   :0.2461   Mean   :0.2679   Mean   :0.2626
##  3rd Qu.: 25.00   3rd Qu.:0.4740   3rd Qu.:0.0000   3rd Qu.:0.0000
##  Max.   :353.00   Max.   :0.7500   Max.   :9.0000   Max.   :9.0000
##       WL%              HOF
##  Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.3850   1st Qu.:0.0000
```

```
##  Median :0.4841   Median :0.0000
##  Mean   :0.4568   Mean   :0.2042
##  3rd Qu.:0.5598   3rd Qu.:0.0000
##  Max.   :1.0000   Max.   :1.0000
```

Overall cleaning and export combined dataset

```
# Make games relative to number of relative season games
df_fb$GR <- df_fb$G / 16
df_bkb$GR <- df_bkb$G / 82
df_bb$GR <- df_bb$G / 162
df_h$GR <- df_h$G / 82
# Make playoff games relative to minimum games needed to win championship
df_fb$PGR <- df_fb$PG / 3
df_bkb$PGR <- df_bkb$PG / (4*4)
df_bb$PGR <- df_bb$PG / (3+4+4)
df_h$PGR <- df_h$PG / (4*4)
# finalize columns and order them
#  Doesn't include YE (last year coaching).
#    Could be useful, but too many values to be a factor
#    and hard to make relative value as sports originated at different times
#  Doesn't include Y as this is highly correlated to GR (games relative)
final_cols <- c('N','GR','WL%','PGR','PWL%','CC','C','HOF')
df_fb_f <- df_fb[final_cols]
df_bkb_f <- df_bkb[final_cols]
df_bb_f <- df_bb[final_cols]
df_h_f <- df_h[final_cols]
# set sport variable
df_fb_f[,'Sport'] <- 'football'
df_bkb_f[,'Sport'] <- 'basketball'
df_bb_f[,'Sport'] <- 'baseball'
df_h_f[,'Sport'] <- 'hockey'
# combine datasets
library(data.table)
df_final <- rbindlist(list(df_fb_f,df_bkb_f,df_bb_f,df_h_f))
df_final <- df_final[order(df_final$N,decreasing=F),]# output to csv file
# account for championships before conference championships
df_final$CC <- pmax(df_final$C,df_final$CC)
# export dataset
write.csv(df_final,'data/all_coaches.csv',row.names=F)
head(df_final)
```

```
##               N       GR       WL%       PGR  PWL% CC C HOF    Sport
## 1:     AJ Hinch 6.308642 0.5580000 4.5454545 0.560  2 1   0 baseball
## 2:  Aaron Boone 2.000000 0.6270000 1.2727273 0.500  0 0   0 baseball
## 3: Aaron Kromer 0.375000 0.3330000 0.0000000 0.000  0 0   0 football
## 4:   Abe Gibron 2.625000 0.2740000 0.0000000 0.000  0 0   0 football
## 5:    Adam Gase 4.000000 0.4690000 0.3333333 0.000  0 0   0 football
## 6:   Adam Oates 1.585366 0.5752212 0.4375000 0.429  0 0   0   hockey
```